Documento de Trabajo 95-04
Serie de Estadística y Econometría 01
Abril 1995

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

# ¿EXISTE UN SESGO DE INACTIVIDAD EN LA ENCUESTA DE POBLACION ACTIVA?

Teresa Villagarcía*

Resumen

En este trabajo se presenta un procedimiento para contrastar si existe un sesgo de inactividad en la Encuesta de Población Activa debido al muestreo por hogares con que se realiza, ya que es más difícil encontrar activos que inactivos en su domicilio en el momento de realización de la encuesta. Para contrastar ésto se estima la tasa de actividad de los varones en edades próximas a la jubilación mediante dos estimadores, uno de ellos insensible al sesgo. Los resultados confirman la existencia del sesgo y puede comprobarse que cumple las propiedades que cabría esperar teóricamente.

Palabras clave
Sesgo de Inactividad; Función Supervivencia.

*Departamento de Estadística y Econometría, Universidad Carlos III de Madrid. Este trabajo es parte de un proyecto más amplio, realizado por investigadores de la Universidad Carlos III de Madrid y financiado por la Fundación Caja de Madrid.

# 1 Introducción y objetivos.

La Encuesta de Población Activa (EPA) es una encuesta transversal que se realiza trimestralmente. Su importancia es grande tanto por su cobertura, representativa de todos los participantes en el mercado laboral, como por su carácter continuo que permite un seguimiento de la información obtenida en la encuesta a través del tiempo.

Los datos obtenidos por la EPA son en general de alta calidad y fiabilidad, como lo demuestra la estabilidad de modelos estimados en EPAs separadas por varios años. Sin embargo, el carácter transversal de la EPA y el tipo de muestreo utilizado pueden generar determinados sesgos, que es conveniente corregir cuando se utilicen sus datos.

Entre los sesgos típicos que cualquier encuesta transversal sufre se encuentra el conocido sesgo de longitud que afecta a las estimaciones de las longitudes del desempleo obtenidas a través de cortes transversales debido a que la probabilidad de captar un período largo de desempleo en un corte transversal es mayor que la de captar uno corto [1].

Otro posible sesgo de la EPA es consecuencia del tipo de muestreo que se utiliza en la recogida de datos. En el caso de la EPA las unidades muestrales son los hogares o viviendas. La recogida de información se realiza por parte de los entrevistadores del Instituto Nacional de Estadística, quienes visitan la sección muestral provistos de la relación de viviendas a visitar y un número de "reservas" para el caso de que sea necesaria alguna sustitución (INE 1987 y 1990). En el momento de la entrevista, las viviendas pueden resultar encuestables o bien ilocalizables, inaccesibles o vacías, en cuyo caso se sustituyen por otras.

El problema que surge de este tipo de muestreo, es que la probabilidad de que una vivienda esté ocupada en el momento de la entrevista no es independiente de la relación con la actividad de sus ocupantes, de modo que es más probable encontrar vacía una vivienda ocupada por una sola persona

[1] Este problema se ha documentado tanto para la EPA (Villagarcía 1988) como para encuestas equivalentes en otros países como es el caso de la Current Population Survey en Estados Unidos (Kiefer et. al 1985)

que trabaja, o en general por grupos humanos con alta tasa de actividad, que viviendas ocupadas por personas o grupos altamente inactivos.

Si esto es así, la encuesta estaría subestimando el número de activos, y dentro de los activos, de ocupados, y sobreestimando el número de inactivos. La importancia de detectar y, en su caso, de cuantificar este problema es evidente en una encuesta cuyos objetivos incluyen la estimación trimestre a trimestre, de la tasa de actividad y de desempleo. Obsérvese que, de existir, el sesgo de inactividad produciría un aumento en el nivel de estas tasas, pero su efecto será casi despreciable en las tasas de variación de las mismas.

La forma tradicional de resolver este problema consiste en realizar con cierta frecuencia una encuesta de control que permita estimar el sesgo y, por tanto, corregir el nivel de las cifras que proporciona la EPA [2]. Sin embargo, en el caso del sesgo de inactividad, esta encuesta de control sería muy costosa de realizar, pues es necesario localizar a los activos que no están en sus hogares durante el horario laboral.

En este artículo se propone un sencillo procedimiento para contrastar estadísticamente la existencia de este sesgo de inactividad y estimarlo, utilizando técnicas estadísticas propias del Análisis de Datos de Supervivencia (ADS). Para ello se estima la tasa de actividad por edades de los varones en edades próximas a la jubilación mediante dos estimadores ligados a dos submuestras distintas de la EPA. El primero parte de la definición tradicional de la tasa de actividad para un grupo poblacional y, si existe el sesgo de inactividad, subestimará la tasa real de actividad por estar subestimando el número de activos. La segunda estimación se realizará a partir de la función de supervivencia de los trabajadores que se retiraron el año anterior a la encuesta. Este dato se conoce porque la EPA del segundo trimestre incluye una pregunta retrospectiva sobre el estado laboral del trabajador un año antes. En este trabajo se demuestra que la función de supervivencia de la edad de retiro estimada mediante esta submuestra es un estimador de la tasa de actividad insensible al sesgo de inactividad, por estar todos los integrantes de la submuestra ya jubilados.

[2] La Current Population Survey, equivalente a la EPA en Estados Unidos, está blindada frente al sesgo de inactividad por dos motivos. En cada hogar se pregunta por los hogares de alrededor, y en segundo lugar, se localiza a los ocupantes en sus trabajos o en horarios no laborales.

A partir de estos dos estimadores puede estimarse el sesgo de inactividad por edades y por tasas de actividad. Es evidente que el sesgo debe ser cero cuando la tasa de actividad de la población sea cero o uno, pues cualquier hogar vacío se sustituirá por otro de las mismas características, y máximo en algún punto intermedio de la tasa de actividad.

El procedimiento tiene dos características interesantes. Por una parte va a utilizar datos de la propia encuesta para detectar y estimar el sesgo. En este sentido el método que se propone es interno y tiene las ventajas obvias de coste. Por otra parte es intuitivo que estudiar la tasa de actividad en edades próximas a la jubilación es altamente informativo, pues en estas edades se producen grandes cambios en la participación laboral de los trabajadores, pasando de tasas de actividad grandes para personas de 50 años a tasas de actividad muy reducidas para los mayores de 65 años.

Se ha estimado el sesgo de inactividad para las encuestas de 1987 a 1991 en función de la edad y de la tasa de actividad. Los resultados obtenidos son totalmente acordes con lo que cabría esperar del carácter del sesgo de inactividad cuyas propiedades teóricas también se han desarrollado.

El esquema del trabajo es el siguiente: En la sección 2 se introducen algunas funciones básicas del Análisis de Datos de Supervivencia. La sección 3 está dedicada a examinar los datos de jubilaciones que capta la EPA y se introducen los dos estimadores cuya comparación va a permitir contrastar la existencia del sesgo de inactividad. La sección 4 presenta los dos estimadores de la tasa de actividad y se estudia el carácter del sesgo y el impacto que tiene sobre los estimadores. La sección 5 muestra los resultados obtenidos para la EPA. Finalmente, la sección 6 está dedicada a las conclusiones.

# 2 Introducción al análisis de datos de supervivencia

El ADS es una técnica estadística que estudia variables aleatorias positivas que representan la duración de procesos como pueden ser la duración de la vida laboral, la edad de jubilación, la duración del desempleo o, en otros

ámbitos no económicos, la vida útil de una bombilla, o el tiempo de remisión de una enfermedad.

Sea $T$ una variable aleatoria positiva que representa la duración de un proceso. Sean $f(t)$ y $F(t)$ las funciones de densidad y distribución de $T$ para un individuo. Se define la función de supervivencia $S(t)$ como

$$S(t_0) = P(T > t_0) = 1 - F(t_0). \tag{1}$$

Si la variable aleatoria es la edad de jubilación de una cohorte de trabajadores, esta función proporciona la probabilidad de que una persona siga activa a la edad $t_0$, y es por tanto una estimación de la tasa de actividad por edades de la cohorte [3].

Se define la tasa de fallos, *Hazard Function*, como

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{2}$$

y representa la probabilidad de que un individuo que sigue activo a determinada edad se jubile inmediatamente. Es importante notar que la tasa de fallos es proporcional a una probabilidad condicionada.

En el análisis de datos de duración es más útil estudiar los datos basándose en las funciones de supervivencia y tasa de fallos que en las habituales de densidad y distribución, ya que la información previa permite optar por distintos modelos. De hecho, los modelos que se utilizan en este ámbito se definen como de Tasa de Fallos creciente, decreciente, constante o en forma de bañera (Lawless 1982) [4]

El problema adicional que se presenta al estudiar variables de duración es que en ocasiones no se observan duraciones completas, sino censuradas. Así,

[3] Se define una cohorte como el conjunto de trabajadores nacidos el mismo año. Se excluyen de las cohortes de trabajadores a los inactivos y a los incapacitados. En el análisis se van a utilizar exclusivamente datos de varones, ya que la participación laboral de las mujeres mayores es muy escasa. Los trabajadores de la misma edad son también una cohorte aproximadamente igual a la anterior.

[4] Así por ejemplo la vida humana tiene una tasa de fallos en forma de bañera: La probabilidad de morir condicionada a estar vivo a la edad $t$, es decreciente desde el nacimiento hasta aproximadamente los cinco años, bastante constante durante la juventud y creciente a partir de los treinta años.

en el caso de las personas activas se sabe que su edad de jubilación, $T$, será superior a su edad actual, $t_c$, pero no su valor exacto. Este tipo de censura donde , $T > t_c$, se denomina censura por la derecha y es muy corriente en variables de duración. La edad actual, $t_c$, se denomina tiempo de censura [5].

En el caso de la jubilación también existe otro tipo de censura, en las personas que ya están jubiladas pero no se conoce su edad de jubilación. En este caso, $T \leq t_c$, y se produce la censura por la izquierda que es muy inhabitual. De hecho no es frecuente encontrar en un problema ambos tipos de censura y ciertas técnicas del Análisis de Datos de Supervivencia como el estimador no-paramétrico de Kaplan Meier (1958) o los modelos de Tasa de fallos proporcional (Cox 1972) no son aplicables.

# 3 Datos

En esta sección se va a analizar la estructura de los datos de edades de jubilación y cómo se captan en una encuesta transversal como la EPA. Estos datos van a permitir definir dos estimadores de la función de supervivencia (que estima la tasa de actividad), uno de ellos sensible y otro insensible al sesgo de inactividad, cuya comparación permitirá detectar la existencia de éste.

La EPA se realiza trimestralmente por el Instituto Nacional de Estadística en 60.000 hogares aproximadamente y recoge datos sobre la situación laboral de los entrevistados. Además, la EPA del segundo trimestre del año recoge datos sobre su situación un año antes, por lo que es posible tener una muestra de trabajadores que se han jubilado durante el último año.

Con los datos de la EPA es posible encontrar varios grupos de interés para el estudio del tránsito a la jubilación:

- *Activos:* Personas que en el momento de la encuesta están activas.

[5] La estimación de la duración del desempleo es un ejemplo clásico de este problema. En datos transversales se sabe que una persona lleva en paro un tiempo $t_c$, pero no se sabe cuanto tiempo más permanecerá desempleada. Kiefer (1988) es una buena introducción a este problema

Constituyen un grupo de datos censurados por la derecha pues su edad de jubilación, $T \geq EDAD$.

- *Jubilados de un Año:* (JA) Personas que están jubiladas en el momento de la encuesta pero que un año antes estaban activas. Su edad de jubilación, $T$, se supone igual a su edad. Aunque puede haber un pequeño error de un año en la edad de jubilación, esto no supone modificaciones en el análisis, pues se puede corregir. Este grupo de personas va a generar la muestra que en la sección 4 se denominará muestra transversal.

- *Jubilados.* Personas que llevan más de un año jubilados. Constituyen un grupo de datos censurados por la izquierda ya que $T < EDAD - 1$.

# 4 Estimadores de la tasa de actividad en edades próximas a la jubilación.

Sea $T$ la variable aleatoria edad de jubilación de las personas de la cohorte nacida en el año $\tau$. Se definen Condiciones de Absoluta Estacionariedad (CAE) cuando se cumple que:

1. Todas las cohortes de trabajadores tienen el mismo tamaño, es decir no consideramos el efecto de la muerte que hace que las cohortes más antiguas sean menores.

2. La distribución de la edad de jubilación permanece invariable entre las cohortes y por tanto el fenómeno de la jubilación permanece invariante, de modo que los trabajadores de 60 años siguen la misma pauta de jubilaciones que los de 64 o los de 65 años.

En estas condiciones, que posteriormente se relajarán, la relación entre los datos que se observan en un corte transversal y las distribuciones de las cohortes se esquematiza en la Figura 1. Las personas de cualquier cohorte van a jubilarse a una edad que se distribuye tal como muestra la figura 1.a. Si,

como muestra el ejemplo, las edades de jubilación varían entre los 63 y los 67 años, de la cohorte de personas nacidas en 1932 que en 1994 tienen 62 años, estarán activos el 100 %, y los primeros que se jubilarán lo harán en 1995 con 63 años. La cohorte de nacidos en 1931, de 63 años, empieza a jubilarse en 1994, la cohorte de 1929, de 64 años, tiene una fracción de personas jubiladas desde hace un año, y otro grupo de personas que se jubilan en 1994.

Nótese que puede estimarse la tasa de actividad de la cohorte de 64 años, como el número de activos de esa edad dividido por el número total de personas de la cohorte de 64 años. Este proceso continúa hasta la cohorte de personas de 67 años cuyos últimos integrantes activos se jubilan en 1994. A partir de ahí, las cohortes están compuestas íntegramente por personas ya jubiladas.

Obsérvese que en el gráfico las partes de las cohortes jubiladas antes de 1994, están a la izquierda del eje vertical del año. Los que se jubilan en 1994 están sobre el eje, y los que siguen activos en 1994 están a la derecha del eje.

La variable $T$, Edad de Jubilación de cualquier individuo de la cohorte $\tau$, representa una V.A. longitudinal, pues para estimar cualquier característica referente al modelo de distribución de probabilidad de $T$ es necesario observar la evolución de la cohorte año tras año. Este tipo de datos permitiría estimar cualquier magnitud relacionada con $T$ que fuera de interés.

Sea $Y$ la V.A. transversal Edad de los jubilados en $t_s$, que en el ejemplo es 1994. Estos datos se obtienen fácilmente en la EPA del segundo cuatrimestre utilizando la muestra JA. En CAE,

$$P(Y = t_j) = P(T = t_j \mid se\ ha\ jubilado\ en\ t_s)$$

$$= \frac{P(T = t_j\ y\ \ se\ ha\ jubilado\ en\ t_s)}{P(\ se\ ha\ jubilado\ en\ t_s)}$$

$$= \frac{P(T = t_j\ y\ Nacido\ en\ t_s - tj)}{P(\ se\ ha\ jubilado\ en\ t_s)} = \frac{P(T = t_j)P(Nacido\ en\ t_s - tj)}{P(\ se\ ha\ jubilado\ en\ t_s)}$$

y por tanto

$$P(Y = t_j) = P(T = t_j)\alpha_j \qquad (3)$$

donde

$$\alpha_j = \frac{P(Nacido\ en\ t_s - tj)}{P(\ se\ ha\ jubilado\ en\ t_s)} \tag{4}$$

En CAE, el tamaño de las cohortes es constante, y coincide con el número de jubilados en cada año por lo que $\alpha_j = 1$ y por tanto $P(Y = t_j) = P(T = t_j)$, es decir la variable aleatoria edad de jubilación longitudinal, $T$ y la variable aleatoria edad de jubilación transversal, $Y$, de los trabajadores que se jubilan un año concreto coinciden y, por tanto, se puede estudiar $T$ a través de $Y$, y concretamente, puede estimarse la función de supervivencia de $T$, $S_T(t)$, a través de la función de supervivencia de $Y$ puesto que, como se acaba de demostrar en la ecuación [ 3], $S_T(t) = S_Y(t)$.

Este resultado es claro en la Figura 1, pues puede comprobarse que la distribución de edades de jubilación de los jubilados en 1994 coincide con la distribución longitudinal de $T$ que muestra la Figura 1.a.

Dado que las variables $Y$ y $T$ tienen una distribución idéntica, puede estimarse su común función de supervivencia a través de dos posibles estimadores:

1. *Estimador Transversal*: Este estimador utiliza la muestra transversal de $Y$ para estimar la función de supervivencia.

$$\hat{S_Y}(t) = \frac{\#Completos\ que\ se\ jubilan\ a\ edad \geq t}{\#Completos} \tag{5}$$

2. *Estimador Longitudinal*: Este estimador utiliza datos por cohortes, ya que aunque no se conoce la evolución de las cohortes, si se conoce parte de esa evolución. Así, de la cohorte que nació en 1929, que en 1994 tiene 65 años, puede estimarse longitudinalmente *un punto de la función de supervivencia* $S_T$, concretamente puede estimarse la probabilidad de seguir trabajando después de los 65 años. Es decir:

$$\tilde{S_T}(t) = \frac{\#Activos\ de\ \ edad\ t}{\#Personas\ de\ \ edad\ t} \tag{6}$$

   Este estimador puede calcularse para todas las cohortes, es decir para todas las edades, obteniéndose una estimación de la función de super-

vivencia que en CAE estimará la común función de supervivencia de $Y$ y $T$.

Estos dos estimadores estiman proporciones y es bien sabido que ambos son centrados y consistentes.

## 4.1 Condiciones no estacionarias

Indudablemente las CAE no se cumplen tal como se han definido. Si se relaja la primera condición manteniendo la segunda, es decir si el tamaño de las cohortes no es el mismo y las cohortes de personas de más edad son menores, entonces $\alpha_i$ definido en la ecuación [ 4], que representa el tamaño de la cohorte respecto al número de jubilados del año, deja de ser igual a uno, y tomará valores mayores que la unidad para las cohortes más recientes y numerosas y valores menores que uno para las cohortes más antiguas y pequeñas.

Admitiendo que el número de personas que componen las cohortes es estrictamente decreciente con la edad, las funciones de supervivencia de $T$ y de $Y$, estarán relacionadas por $S_Y(t) \leq S_T(t)$ verificándose la igualdad cuando $S_Y(t) = S_T(t) = 1$ o $S_Y(t) = S_T(t) = 0$.

En estas condiciones, el estimador longitudinal dado por la ecuación [ 6] seguirá siendo un estimador centrado de $S_T(t)$ si se supone que la muerte afecta por igual a jubilados y activos de la misma edad, mientras que el transversal dado por [ 5] presentará un sesgo a la hora de estimar $S_T(t)$, pues las distribuciones de $Y$ y de $T$ ya no coinciden.

Las Figuras 2 muestran este resultado para datos simulados. Se ha simulado inicialmente una distribución de edades de jubilación que es la que muestra el histograma de la Figura 2.a. A continuación se han supuesto CAE y se ha ido contabilizando el número de activos y jubilados de cada edad en 1994 suponiendo que año a año se repite exactamente la pauta de jubilaciones de la Figura 2.a, y que el número de personas de cada cohorte es el mismo.

La figura 2.b muestra los estimadores transversal y longitudinal calculados para estos datos. Lógicamente ambos coinciden. La figura 2.c muestra los mismos estimadores cuando se relaja la primera CAE, es decir cuando el tamaño de las cohortes no es el mismo. Se ha supuesto un tamaño de cohortes que disminuye linealmente como indica la línea de puntos, es decir que la cohorte de 73 años es la mitad que la de 56 años. En este caso, el estimador longitudinal sigue siendo centrado, pero no el transversal, que es menor.

La segunda CAE es menos importante que la primera, ya que representa la igualdad de la distribución de edades de jubilación de las cohortes. A corto plazo, es previsible que no haya grandes cambios en este fenómeno. En el caso de la Encuesta de Población Activa como se indicará en la sección 5 puede admitirse el cumplimiento de la segunda CAE.

En cualquier caso, el impacto de relajar la segunda Condición de Absoluta Estacionariedad, es que ya no existe una sola variable $T$ común a todas las cohortes, sino una para cada cohorte, $T_\tau$. La V.A. $Y$ es una mezcla de las distribuciones de edades de jubilación $T_\tau$ de las cohortes ponderada por el tamaño de las mismas.

## 4.2 Impacto del sesgo de inactividad sobre los estimadores.

El muestreo de la EPA se realiza sobre hogares. Este tipo de muestreo no es independiente de las variables relacionadas con la actividad, ya que la probabilidad de que un hogar habitado por personas activas y ocupadas esté vacío cuando el entrevistador acude a realizar la encuesta, es mayor que la de un hogar en el que haya personas inactivas.

Si ésto es así, la muestra estará subestimando el número de activos y sobreestimando el número de inactivos, que serán más fáciles de encontrar en sus casas en horario laboral.

El sesgo de inactividad no afecta al estimador transversal definido en la ecuación [ 5], ya que *todos los integrantes de la muestra son jubilados* y por tanto si existe el sesgo lo que hará es aumentar el tamaño muestral y

consecuentemente disminuir el error de estimación del estimador transversal *que, en CAE, seguirá siendo un estimador centrado de $S_T(t)$, o manteniendo las propiedades que le correspondan si las CAE se cumplen parcialmente.*

El estimador longitudinal -ecuación [ 6]- se verá afectado por el sesgo, ya que utiliza para su cómputo tanto activos como jubilados. Si el número de activos está disminuido, el estimador longitudinal subestimará la tasa de actividad de las personas de edad avanzada.

La figura 2.d presenta los dos estimadores, longitudinal y transversal, en CAE pero habiendo disminuido el número de activos para los datos simulados. Como puede comprobarse, el estimador longitudinal es menor que el transversal que sigue siendo centrado.

La tabla 1 resume las características de los dos estimadores en las diversas condiciones que se han estudiado.

Básicamente pueden admitirse las siguientes situaciones:

1. CAE. En este caso los estimadores longitudinal y transversal son centrados para $S_T(t)$.

2. Efecto muerte. En este caso, el estimador longitudinal sigue siendo centrado para $S_T(t)$, pero el transversal ya no lo es y subestima $S_T(t)$. Cabrá esperar entonces que $\hat{S}_Y(t) \leq \tilde{S}_T(t)$.

3. Sesgo de inactividad y CAE. En este caso el estimador transversal sigue siendo centrado para $S_T(t)$. El estimador longitudinal subestimará $S_T(t)$. Nótese que la dirección de variación en las condiciones [2] y [3] es inversa, cabe esperar por tanto que $\hat{S}_Y(t) \geq \tilde{S}_T(t)$.

4. Sesgo de inactividad y no CAE. Combinación de las situaciones [2] y [3]. Si el sesgo de inactividad predomina puede llegarse a $\hat{S}_Y(t) \geq \tilde{S}_T(t)$.

Tabla 1

| | Estimador | |
|---|---|---|
| | Longitudinal $\tilde{S}_T(t)$ | Transversal $\hat{S}_Y(t)$ |
| Datos | Cohortes | Jubilados el año anterior |
| Expresión | $\tilde{S}_T(t) = \frac{\#Activos\ de\ edad\ t}{\#Personas\ de\ edad\ t}$ | $\hat{S}_Y(t) = \frac{\#Completos\ que\ se\ jubilan\ a\ edad>t}{\#Completos}$ |
| 1. CAE | $S_T(t) = S_Y(t)$ Centrado | $S_T(t) = S_Y(t)$ Centrado |
| 2. Varía el tamaño de la cohorte | $S_T(t) \geq S_Y(t)$ Centrado | Sesgado como estimador de $S_T(t)$ |
| 3. Distribución de $T$ varía | Existe una V.A. $T_\tau$ para cada cohorte $\tau$ | $Y$ es mezcla de las distribuciones $T_\tau$ |
| 4. CAE y Sesgo | Sesgado. Subestima $S_T(t) = S_Y(t)$ | $S_T(t) = S_Y(t)$ Centrado |
| 5. Sesgo y tamaño de cohorte variable | Mezcla de los casos 2 y 4 | Mezcla de los casos 2 y 4 |

## 4.3 Carácter del sesgo.

El sesgo de inactividad tiene un carácter semejante a la tasa de propagación de las enfermedades contagiosas [6]: es nulo cuando la tasa de actividad es cero o uno, ya que cualquier hogar sustituido tendrá el mismo tipo de relación con la actividad que el vacío, y será máximo para alguna tasa de actividad intermedia.

Suponiendo que $t_a$ sea la tasa de actividad de la población, que $p_a$ y $p_i$ sean las probabilidades de que un hogar de activos y un hogar de inactivos estén ocupados cuando llega el entrevistador, la tasa de actividad que estimaría la muestra será:

$$t_m = \frac{p_a t_a}{p_a t_a + (1 - t_a) p_i} \tag{7}$$

[6]La tasa de propagación de enfermedades contagiosas es baja para proporciones muy bajas y muy altas de infectados, y es máxima en algún punto intermedio

y el sesgo

$$sesgo = t_a - t_m = \frac{(t^2 - t)d}{td + p_i} \tag{8}$$

donde $d = p_a - p_i$. Suponiendo que $p_a \leq p_i$, entonces $t_m \leq t_a$. Obviamente el sesgo nulo se obtiene para d=0, es decir cuando no existe diferencia entre el comportamiento entre activos e inactivos. El máximo sesgo se produce cuando la tasa de actividad sea:

$$t_a = (-p_i + \sqrt{p_a p_i})/d \tag{9}$$

Es por tanto esperable encontrar un sesgo que presente un máximo en función de la tasa de actividad de la población.

# 5 Resultados: aplicación a la EPA

Como se ha indicado en la sección 3, la EPA del segundo trimestre permite obtener una muestra de los trabajadores que se han jubilado en el último año. Esta muestra transversal, si las condiciones son estacionarias, permitiría reconstruir la distribución longitudinal de edades de jubilación.

La primera CAE evidentemente no se cumple. El Gráfico 3 muestra el número total de varones activos y jubilados por edades registrados en las encuestas entre 1987 y 1991. Los datos están normalizados y suavizados mediante una media móvil con pesos decrecientes respecto al valor central. El cambio de curvatura en torno a los 65 años se debe, posiblemente, al sesgo de inactividad, ya que a los 65 años prácticamente todos los hombres se han jubilado y por tanto la posibilidad de encontrarlos es mayor.

La segunda CAE se puede considerar que se cumple. En el apéndice se presentan cinco modelos Weibull para explicar la edad de jubilación ajustados a los jubilados completos JA, tomados de Villagarcía (1995). Como puede comprobarse los coeficientes de los modelos para los cinco años pueden pasar un contraste múltiple de igualdad, con excepción de los estudios en 1990, por lo que puede imponerse que la distribución de $Y \mid X$ permanece

constante durante el período 1987-1991. Consecuentemente, la distribución longitudinal de $T \mid X$ también permanecerá constante, siendo $X$ un vector de variables explicativas.

Este resultado es lógico, pues entre 1987 y 1991 no ha habido cambios significativos ni en la legislación concerniente a jubilaciones, ni en la metodología de la encuesta. En estas condiciones cabe esperar, que si no hubiera sesgo de inactividad el estimador transversal $\hat{S}_Y$ obtenido a través de la muestra de jubilados del año sea menor que el estimador longitudinal $\tilde{S}_T$ obtenido de las cohortes. Este descenso del estimador transversal es debido al efecto de la mortalidad. La situación se corresponde con la figura 2.c de las simulaciones.

La figura 4 muestra los estimadores longitudinal y transversal para 1990 y 1991 respectivamente [7]. Es interesante notar la estabilidad de cada una de las funciones estimadas que, tal como muestra la figura, son prácticamente indistinguibles y pasan un contraste de igualdad, de manera que los estimadores longitudinales por una parte y los transversales por otra, pueden considerarse iguales para los dos años. No se han representado las bandas de confianza por ser muy estrechas, ya que la muestra es grande y las varianzas de los estimadores corresponden a las de una proporción, $p(1-p)/n$. Sin embargo, es importante indicar que a partir de los 65 años, o para tasas de actividad menores de 0.2, ambos estimadores, transversal y longitudinal, están dentro de las bandas de confianza, por lo que el sesgo puede considerarse cero a partir de esas tasas de actividad.

En la Figura 4, puede observarse que el estimador longitudinal para los dos años es menor que el transversal. Este efecto, que corresponde a la figura 2.d de las simulaciones, indica la existencia de un sesgo lo suficientemente fuerte como para paliar el de mortalidad que tendería a situar el estimador transversal por debajo del longitudinal.

Como se ha indicado, el estimador transversal se ve afectado a la baja por la mortalidad, pero no se ve afectado por el sesgo de inactividad. Por el contrario, el estimador longitudinal se ve afectado por el sesgo pero no por la mortalidad. El efecto muerte puede estimarse, por tanto, partiendo del esti-

[7] El efecto se ha estudiado para los años 1987-1991. En todos los años aparece el sesgo con claridad, aunque como indica la Tabla 2, para el período 1987-89 los sesgos son algo menores que para el período 1990-91.

mador transversal y utilizando la ecuación [4] con los valores estimados de $\hat{\alpha}_j$ y $P(\widehat{Y = t_j})$ . Con estos valores puede calcularse la función de supervivencia corregida del sesgo de la mortalidad $\hat{S}_Y^*$. Hay que resaltar que la diferencia entre $\hat{S}_Y^*$ y $\hat{S}_Y$ es muy pequeña, como puede observarse en la Figura 2.c de las simulaciones. En el caso de la EPA los valores $\hat{S}_Y^*$ y $\hat{S}_Y$ apenas difieren.

La figura 5 muestra los sesgos estimados y suavizados para 1990 y 1991 en función de la tasa de actividad corregida del efecto mortalidad, $\hat{S}_Y^*$, suavizada. Como puede observarse, y era esperable según la ecuación [ 9], el sesgo presenta un máximo, que se produce para una tasa de actividad entre 0.6 y 0.7. La diferencia entre los sesgos estimados en los dos años es muy pequeña. La lectura de la figura 5 es interesante: Si la tasa de actividad de la población masculina fuera de 0.7, correspondiente al valor máximo del sesgo, la tasa de actividad que estimaría la EPA sería, según la ecuación [ 8], de aproximadamente 0.6, pues el sesgo es prácticamente 0.1. En el caso español, la tasa de actividad masculina estimada en 1991 es del 65.8% por lo que la tasa de actividad real, estaría en torno al 75% que es una cifra más acorde a las obtenidas en los países de nuestro entorno [8].

Se ha estimado un modelo polinómico para obtener una expresión analítica del sesgo con los datos de 1990 y 1991 [9]:

$$\begin{array}{lcccc} SESGO = & -0.007 & +0.16t_a & +0.15t_a^2 & -0.24t_a^3 \\ & (-5.22) & (6.73) & (2.3) & (-5.66) \\ & & & & R^2 = 98.55 \end{array} \qquad (10)$$

La Tabla 2 presenta los valores estimados del sesgo y de la tasa de actividad corregida del efecto mortalidad por edades para los años 1987 a 1991. Los asteriscos indican que las curvas de supervivencia longitudinal y transversal

[8] Los sesgos se han estimado exclusivamente para hombres, ya que la población femenina jubilada el año anterior a la encuesta es muy pequeña. En estas condiciones, la generalización del valor estimado del sesgo a la población total, implica aceptar la hipótesis de que el comportamiento de hombres y mujeres es el mismo en cuanto a sus probabilidades de estar en su casa siendo activos o inactivos. Esta hipótesis es indudablemente fuerte, pero si se admite, puede calcularse la tasa de actividad corregida ponderando la tasa masculina y la femenina y corrigiendo ambas.

[9] Entre paréntesis estadístico t

pasan un contraste de igualdad para esos valores. Esto no quiere decir que no haya sesgo para las tasas de actividad más bajas, sino que éste es lo suficientemente pequeño para que se confunda con el error de estimación. De nuevo esta evidencia es acorde con la teoría, pues el sesgo debe ser pequeño para tasas de actividad bajas. La Tabla 3 presenta los valores del sesgo en función de la tasa de actividad según el modelo estimado en la ecuación [ 10].

Finalmente, es importante indicar que en la EPA existe información sobre el número de viviendas vacías que se han encontrado en el muestreo (INE 1987). En 1987, por ejemplo, se registra un promedio entre provincias del 14.26% de viviendas vacías, siendo el promedio para todas las viviendas encuestadas del 13,8%. Un estudio de la proporción de viviendas vacías por provincias, indica que es mayor para las provincias más turísticas y para las que han sufrido mucha emigración, lo que es muy razonable, pues es lógico encontrar un mayor parque de viviendas vacías en ambos casos.

El sesgo estimado, alcanza valores menores que los aquí referidos, como es de esperar, pues evidentemente, aunque en ocasiones una vivienda esté vacía debido al sesgo, en otros casos la vivienda estará vacía. Este contraste externo es otra evidencia a favor del modelo propuesto.

# 6 Conclusiones

En este trabajo se ha contrastado la existencia del sesgo de inactividad en la EPA. Este sesgo tiene su origen en el muestreo por hogares con que se obtienen los datos de la EPA. Los hogares con altas tasas de actividad, es más probable que estén vacíos en el momento de la primera entrevista, en cuyo caso la vivienda se sustituye por otra. Si esto es así, la tasa de actividad de la encuesta estará subestimando la auténtica tasa de actividad de la población. Este problema en principio no afecta a la encuesta equivalente en Estados Unidos, la Current Population Survey, ya que en cada hogar se pregunta por los de alrededor y se localiza a los ausentes en sus lugares de trabajo.

El contraste de existencia del sesgo de inactividad se basa en la comparación de dos estimadores de la función de supervivencia para trabajadores en edades próximas a la jubilación. De estos dos estimadores, el transversal

es insensible al sesgo por estar calculado a partir de personas inactivas. El estimador longitudinal se ve afectado por el sesgo y la tasa de actividad muestral subestimará la de la población. La comparación de ambos estimadores permite poner de manifiesto la existencia del sesgo y cuantificarlo. En el caso español, con tasas de actividad masculinas muestrales en 1991 de 65.8%, el sesgo se situaría en torno a 9 puntos porcentuales, proporcionando una tasa de actividad de la población masculina del 74%.

Se ha demostrado la dependencia del sesgo de la tasa de actividad, siendo cero el sesgo para tasas de actividad cero o uno, y obteniéndose un máximo para algún valor intermedio que depende de la probabilidad promedio de estar ocupadas las viviendas con alta y baja tasa de actividad. Las estimaciones del sesgo obtenidas son compatibles con este comportamiento del sesgo. Es importante resaltar que el sesgo de inactividad afecta fuertemente al nivel de la tasa de actividad, pero de manera mucho más suave a su variación interanual, de modo que los cambios en la tasa de actividad pueden considerarse centrados.

Destacaremos dos importantes extensiones en este trabajo. En primer lugar el sesgo de inactividad puede afectar a otras magnitudes además de la tasa de actividad. Así por ejemplo parece lógico considerar que dentro del grupo de activos de cualquier edad, los parados tendrán un comportamiento más parecido a los inactivos que los ocupados, de modo que la probabilidad de ocupación de un hogar de parados será mayor que la de un hogar de ocupados. Si esto es así, la tasa de desempleo en España estará sobreestimada debido a una *subestimación del número de ocupados.* Los datos de las tablas 2 y 3 pueden ser utilizados para valorar en qué medida está subestimada la tasa de desempleo sin más que sustituir la tasa de actividad por la tasa de ocupación. Indudablemente para considerar adecuada esta extensión, hay que admitir que el comportamiento de ocupados y parados es semejante al de activos e inactivos, pues como indica la ecuación [ 8] el sesgo estaría parametrizado en este caso por las probabilidades de ocupación de hogares de parados y empleados. El problema podría resolverse con una batería de preguntas retrospectivas en la misma línea de las que permiten reconstruir la submuestra de jubilados del año, aunque existen dos problemas adicionales, al no poder admitir la segunda CAE, ya que la dependencia de la duración del desempleo respecto del ciclo económico es clara y tener que corregir el

efecto del sesgo de longitud para las observaciones censuradas.

En segundo lugar, hay que indicar que la EPA es una encuesta robusta en el muestreo, es decir que los resultados que se obtienen para un año se repiten en años sucesivos. Esto permite realizar análisis de estabilidad de los resultados y determinar pautas para corregir los sesgos que podrían utilizarse en años sucesivos.

Apéndice: Modelos Weibull estimados para jubilados completos JA.

| VARIABLE | 1987 | 1988 | 1989 | 1990 | 1991 |
|---|---|---|---|---|---|
| CONSTANTE | 4,2 | 4,19 | 4,19 | 4,19 | 4,19 |
| | (0,007) | (0,007) | (0,006) | (0,007) | (0,006) |
| AES | -0,004 | -0,003 | -0,003 | -0,0013 | -0,004 |
| | (0,0018) | (0,0018) | (0,0016) | (0,0019) | (0,0017) |
| AES2 | 0,0003 | 0,00036 | 0,0003 | 0,0001 | 0,0003 |
| | (0,0001) | (0,0001) | (0,0001) | (0,0001) | (0,0001) |
| ASALPUB | -0,025 | -0,034 | -0,03 | -0,04 | -0.02 |
| | (0,005) | (0,006) | (0,005) | (0,006) | (0,005) |
| ASALPRI | -0,020 | -0,014 | -0,025 | -0,018 | -0,013 |
| | (0,005) | (0,005) | (0,005) | (0,005) | (0,0045) |
| CONSTR | -0,015 | -0,015 | -0,016 | -0,02 | -0,02 |
| | (0,007) | (0,008) | (0,007) | (0,009) | (0,007) |
| INDUS | -0,018 | -0,02 | -0,03 | -0,01 | -0,02 |
| | (0,005) | (0,005) | (0,005) | (0,005) | (0,005) |
| PARADO | -0,05 | -0,036 | -0,054 | -0,048 | -0,04 |
| | (0,009) | (0,009) | (0,008) | (0,009) | (0,009) |
| ESCALA | 0,041 | 0,049 | 0,045 | 0,047 | 0,043 |
| | (0,0013) | (0,0013) | (0,0013) | (0,0014) | (0,0013) |
| Número de Observaciones | 532 | 595 | 608 | 578 | 550 |

Variable explicada: Edad de Jubilación.
Entre paréntesis errores estándar.

**Donde:**

**AES** Años de estudio.

**AES2** Años de estudio al cuadrado.

**ASALPUB** Variable dicotómica. Vale 1 para asalariados públicos.

**ASALPRI** Variable dicotómica. Vale 1 para asalariados privados.

**CONSTR** Variable dicotómica. Vale 1 para trabajadores del sector construcción.

**INDUS** Variable dicotómica. Vale 1 para trabajadores del sector industria.

**PARADO** Variable dicotómica. Vale 1 para jubilados parados un año antes.

# Tabla 2 SESGOS EN FUNCIóN DE LAS TASAS DE ACTIVIDAD Y LA EDAD

| Edad | 1987 | | 1988 | | 1989 | | 1990 | | 1991 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{S}_Y^*$ | $\hat{S}_Y^* - \tilde{S}_T$ | $\hat{S}_Y^*$ | $\hat{S}_Y^* - \tilde{S}_T$ | $\hat{S}_Y^*$ | $\hat{S}_Y^* - \tilde{S}_T$ | $\hat{S}_Y^*$ | $\hat{S}_Y^* - \tilde{S}_T$ | $\hat{S}_Y^*$ | $\hat{S}_Y^* - \tilde{S}_T$ |
| 55 | 0.99 | 0.07 | 0.97 | 0.04 | 0.97 | 0.08 | 0.98 | 0.07 | 0.99 | 0.06 |
| 56 | 0.97 | 0.08 | 0.94 | 0.06 | 0.94 | 0.05 | 0.96 | 0.08 | 0.97 | 0.07 |
| 57 | 0.95 | 0.08 | 0.90 | 0.05 | 0.91 | 0.05 | 0.95 | 0.07 | 0.94 | 0.09 |
| 58 | 0.93 | 0.06 | 0.87 | 0.04 | 0.87 | 0.05 | 0.91 | 0.07 | 0.91 | 0.07 |
| 59 | 0.89 | 0.07 | 0.84 | 0.06 | 0.85 | 0.04 | 0.88 | 0.08 | 0.88 | 0.06 |
| 60 | 0.78 | 0.07 | 0.76 | 0.04 | 0.73 | 0.05 | 0.77 | 0.10 | 0.77 | 0.10 |
| 61 | 0.71 | 0.06 | 0.68 | 0.07 | 0.67 | 0.05 | 0.70 | 0.10 | 0.71 | 0.11 |
| 62 | 0.64 | 0.02 | 0.60 | 0.03 | 0.63 | 0.08 | 0.65 | 0.10 | 0.64 | 0.09 |
| 63 | 0.60 | 0.05 | 0.56 | 0 | 0.57 | 0.05 | 0.59 | 0.12 | 0.60 | 0.10 |
| 64 | 0.55 | 0.06 | 0.51 | 0.04 | 0.48 | 0.01 | 0.54 | 0.07 | 0.55 | 0.12 |
| 65 | 0.14 | * | 0.12 | * | 0.13 | * | 0.12 | * | 0.16 | 0.03 |
| 66 | 0.07 | * | 0.04 | * | 0.07 | * | 0.06 | * | 0.07 | * |
| 67 | 0.05 | * | 0.04 | * | 0.06 | * | 0.04 | * | 0.04 | * |

Los asteriscos indican sesgos que los estimadores longitudinal y transversal pasan un contraste de igualdad, por lo que el sesgo estimado es menor que el error de la estimación.

Tabla 3 SESGO ESTIMADO EN FUNCIóN DE LA TASA DE ACTIVIDAD

| Tasa de Actividad | Sesgo |
|---|---|
| 0.05 | 0.0013 |
| 0.10 | 0.0103 |
| 0.15 | 0.0196 |
| 0.20 | 0.0291 |
| 0.25 | 0.0386 |
| 0.30 | 0.0480 |
| 0.35 | 0.0571 |
| 0.40 | 0.0656 |
| 0.45 | 0.0735 |
| 0.50 | 0.0805 |
| 0.55 | 0.0864 |
| 0.60 | 0.0912 |
| 0.65 | 0.0945 |
| 0.70 | 0.0962 |
| 0.75 | 0.0961 |
| 0.80 | 0.0941 |
| 0.85 | 0.0900 |
| 0.90 | 0.0835 |
| 0.95 | 0.0746 |

# Referencias.

Cox D.R. (1972) Regression Models and Life Tables (with discussion).*J.R. Stat. Soc. B, 34.*

INE (1987) "Incidencias en los trabajos de campo".

INE (1990) "Evaluación de la calidad de los datos en la Encuesta de Población Activa".

Kaplan E.L. y Meier P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* 53, 457-481.

Kiefer N.M. (1988) Economic Duration Data and Hazard Functions. *Journal of Economic Literature. Vol XXVI.*

Lawless J.F. (1982) *Statistical Models and Methods for Lifetime Data.* Wiley.

Villagarcía, T. (1988) "Estimación de la distribución de duraciones de desempleo en España: 1976-1984". *Estadística Española. Vol. 30 Num. 119. pp 445-470.*

Villagarcía, T. (1995) "Análisis econométrico del tránsito a la jubilación para trabajadores de edad avanzada" *Investigaciones Económicas. Enero.*

# Figura 1.

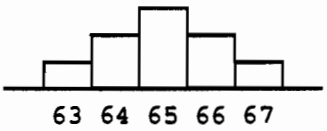


Figura 1.a

Distribución de edades de jubilación común a todas las cohortes.

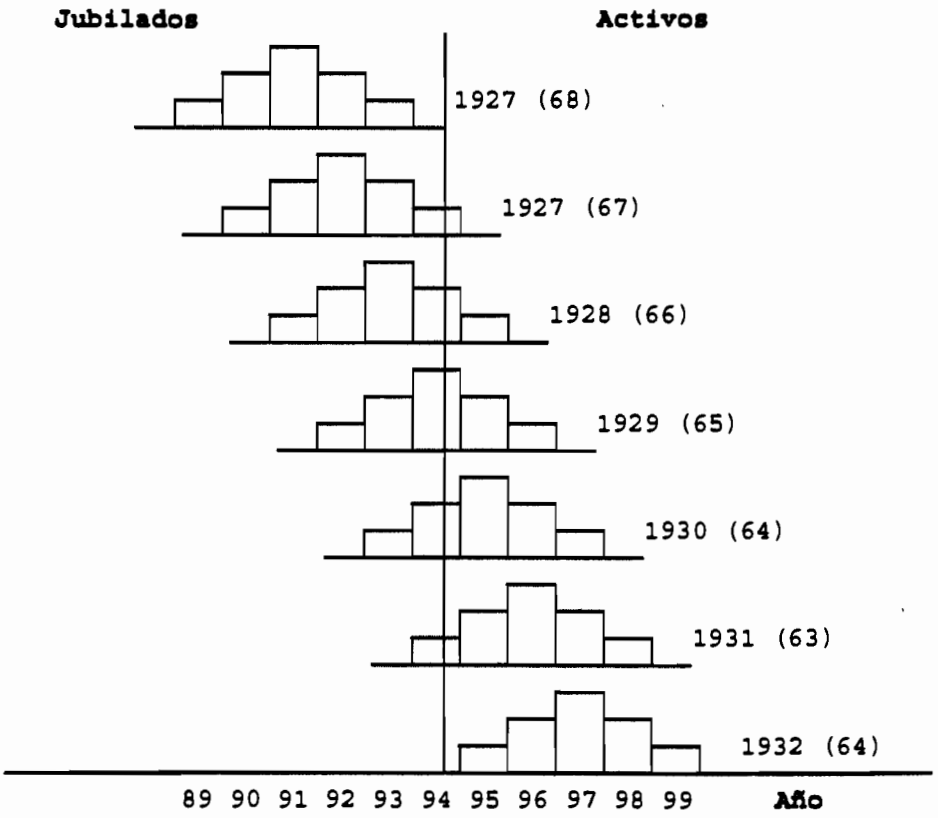


Figura 1.b

Evolución de las cohortes de trabajadores y forma de captarlas de una encuesta transversal.

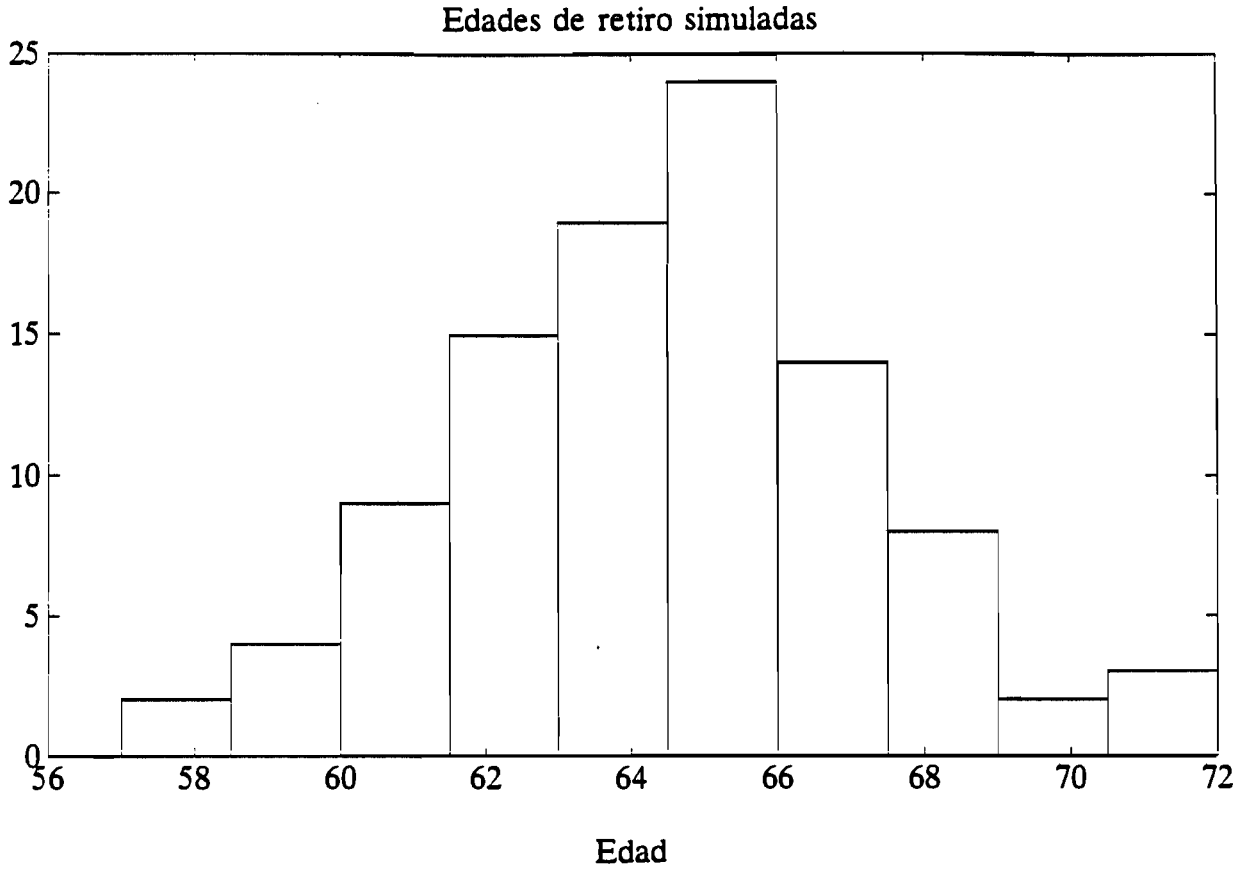
Edades de retiro simuladas
25
20
15
10
5
0
56
58
60
62
64
66
68
70
72
Edad

Figura 2.a

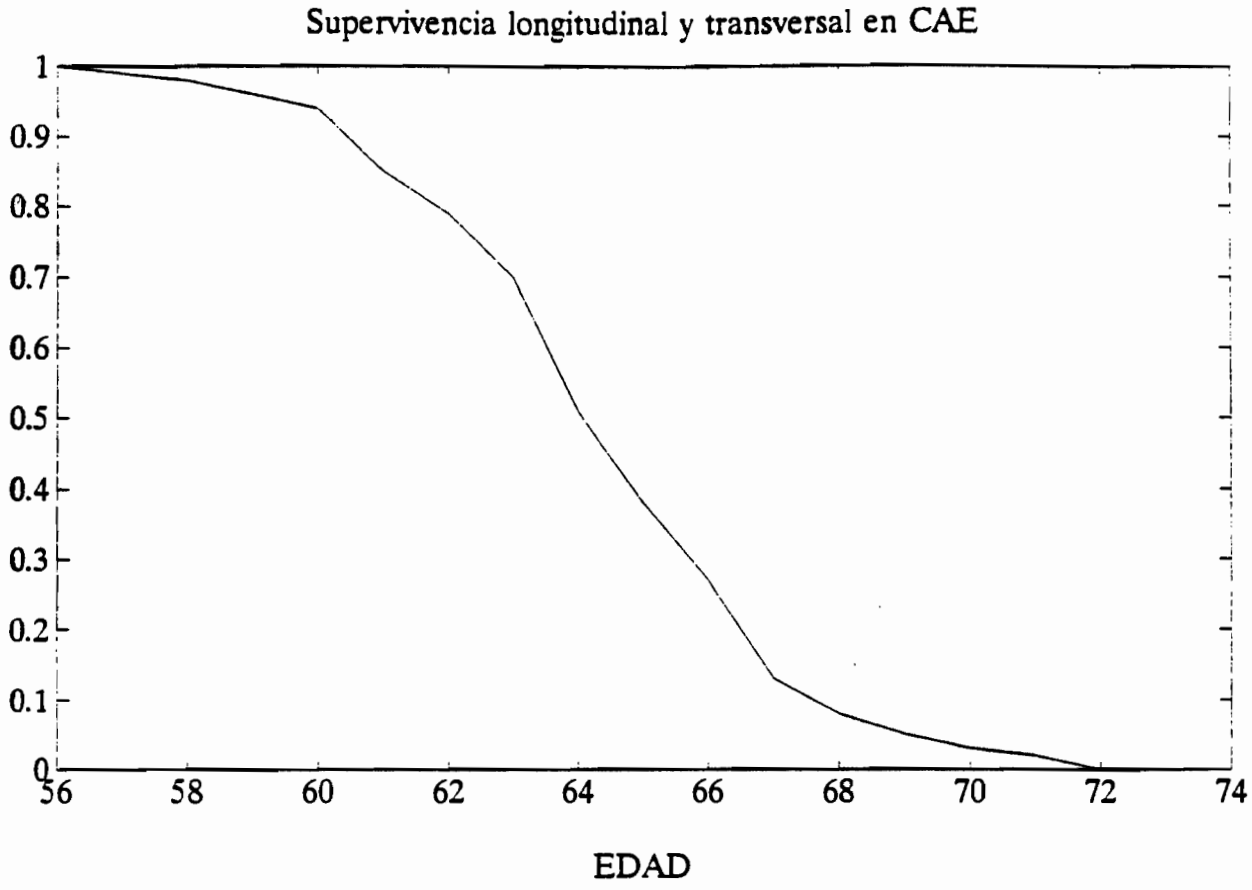
Supervivencia longitudinal y transversal en CAE
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
56
58
60
62
64
66
68
70
72
74
EDAD

Figura 2.b

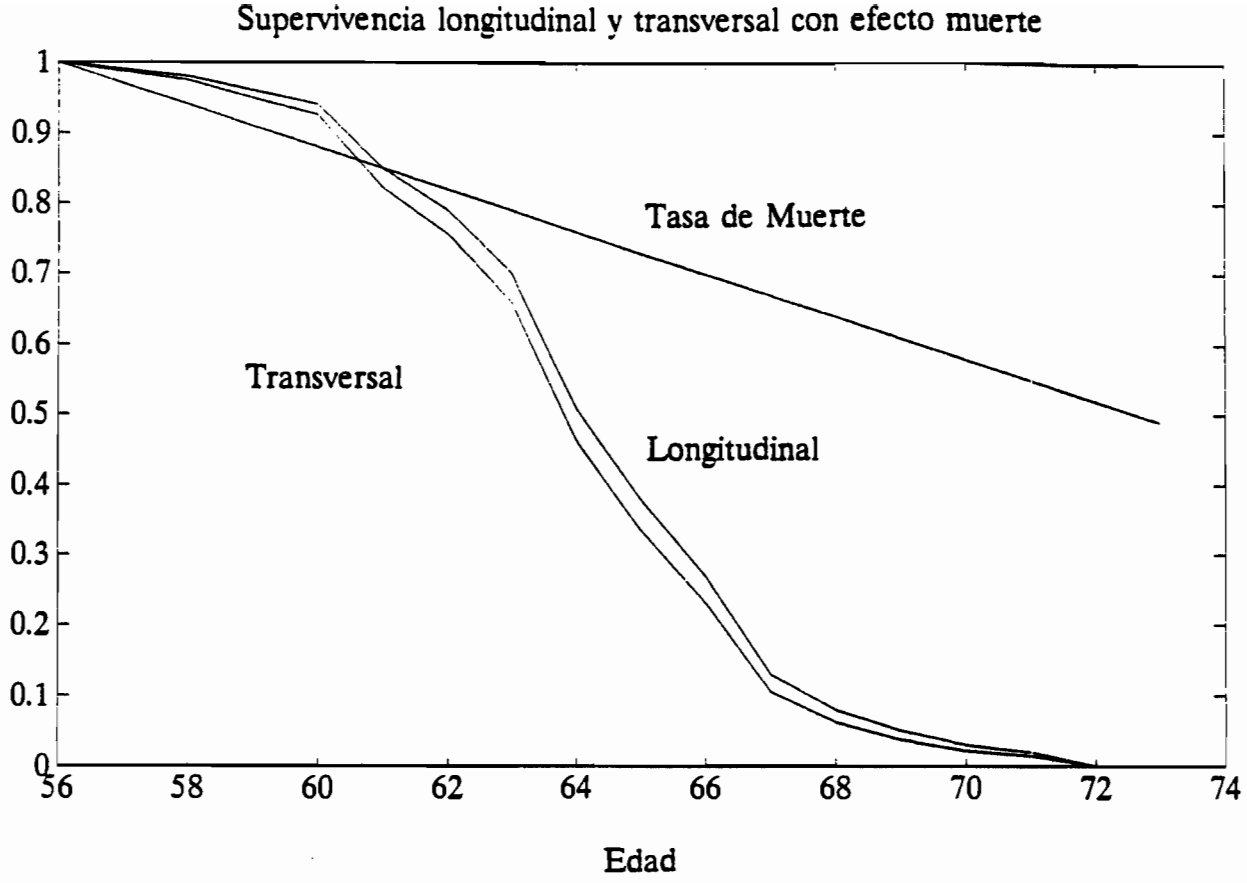
Supervivencia longitudinal y transversal con efecto muerte
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
56
58
60
62
64
66
68
70
72
74
Tasa de Muerte
Transversal
Longitudinal
Edad

Figura 2.c

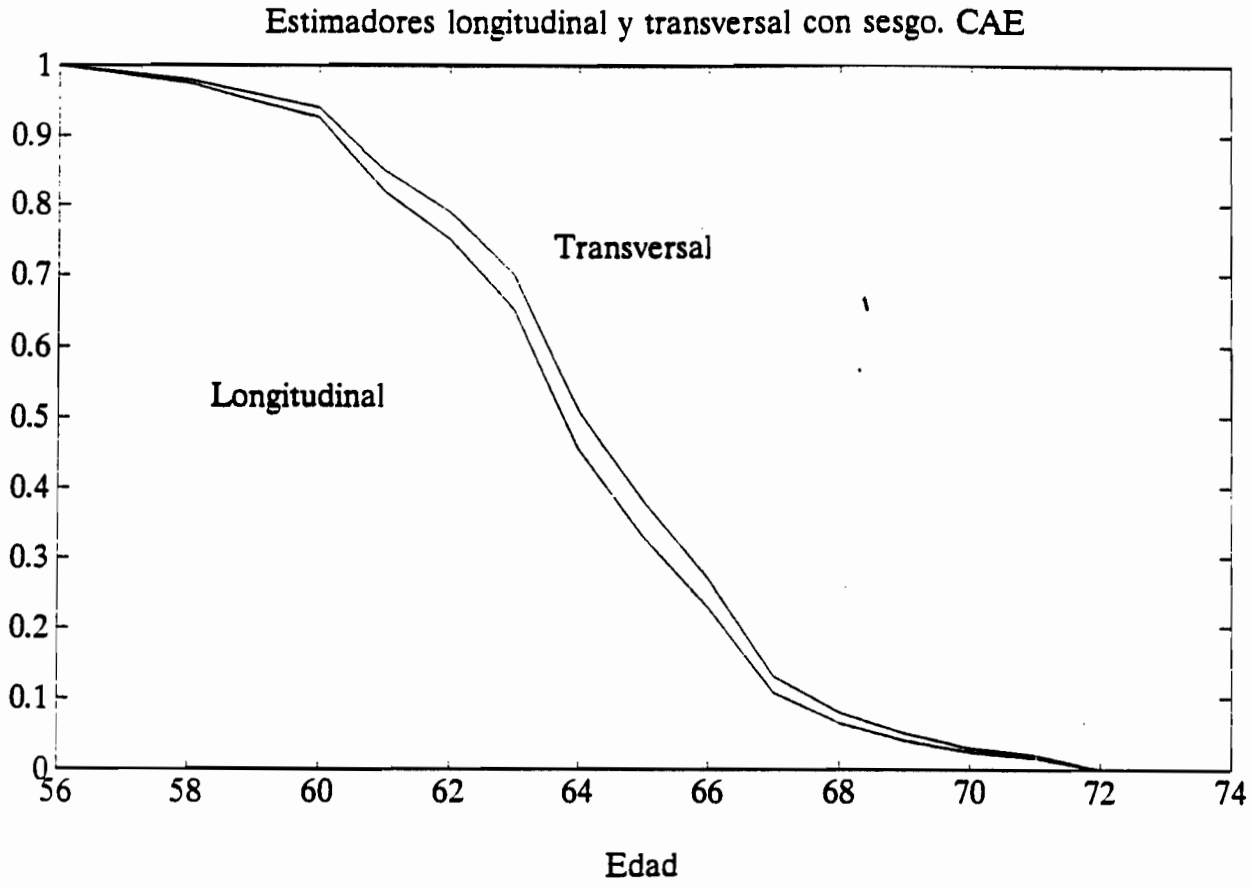
Estimadores longitudinal y transversal con sesgo. CAE
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
56
58
60
62
64
66
68
70
72
74
Transversal
Longitudinal
Edad

Figura 2.d

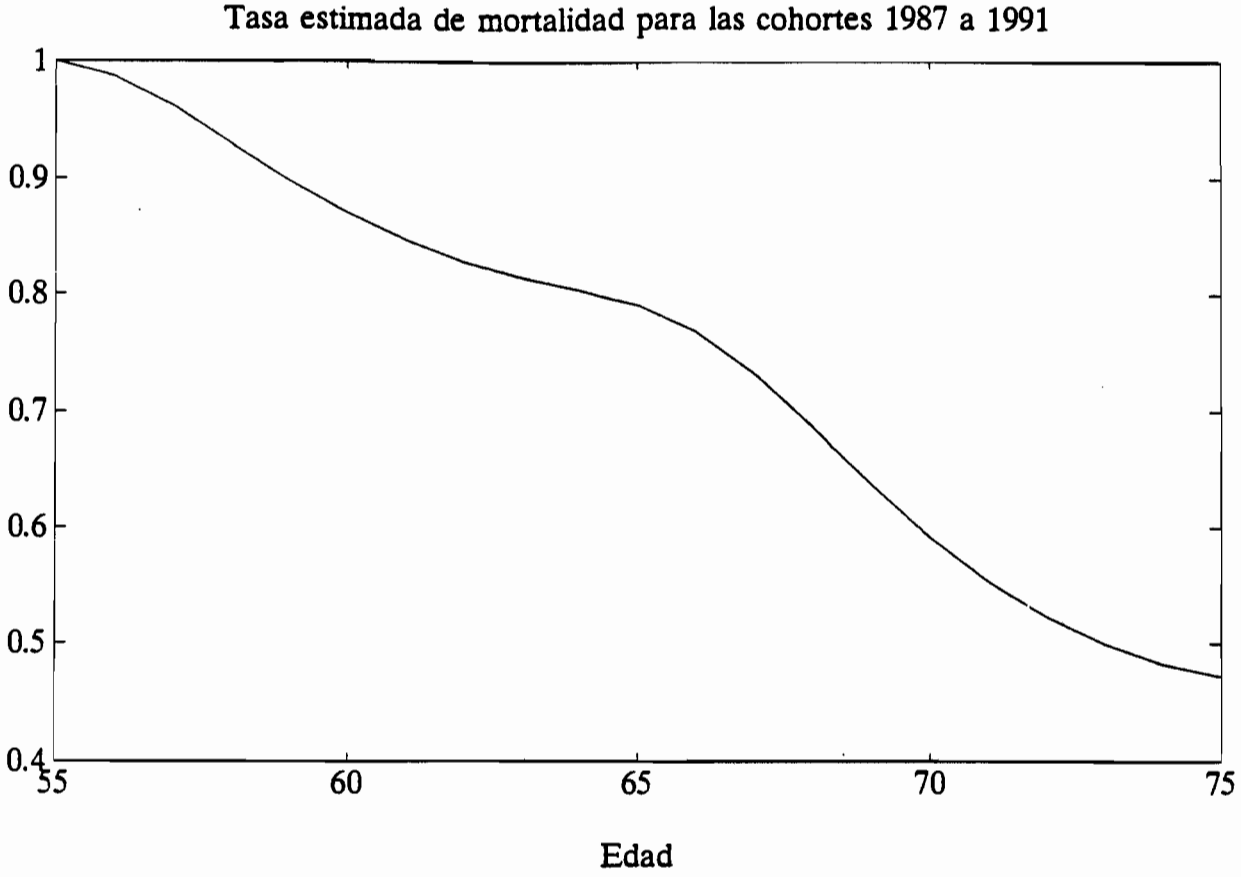
Tasa estimada de mortalidad para las cohortes 1987 a 1991
1
0.9
0.8
0.7
0.6
0.5
0.4
55
60
65
70
75
Edad

Figura 3

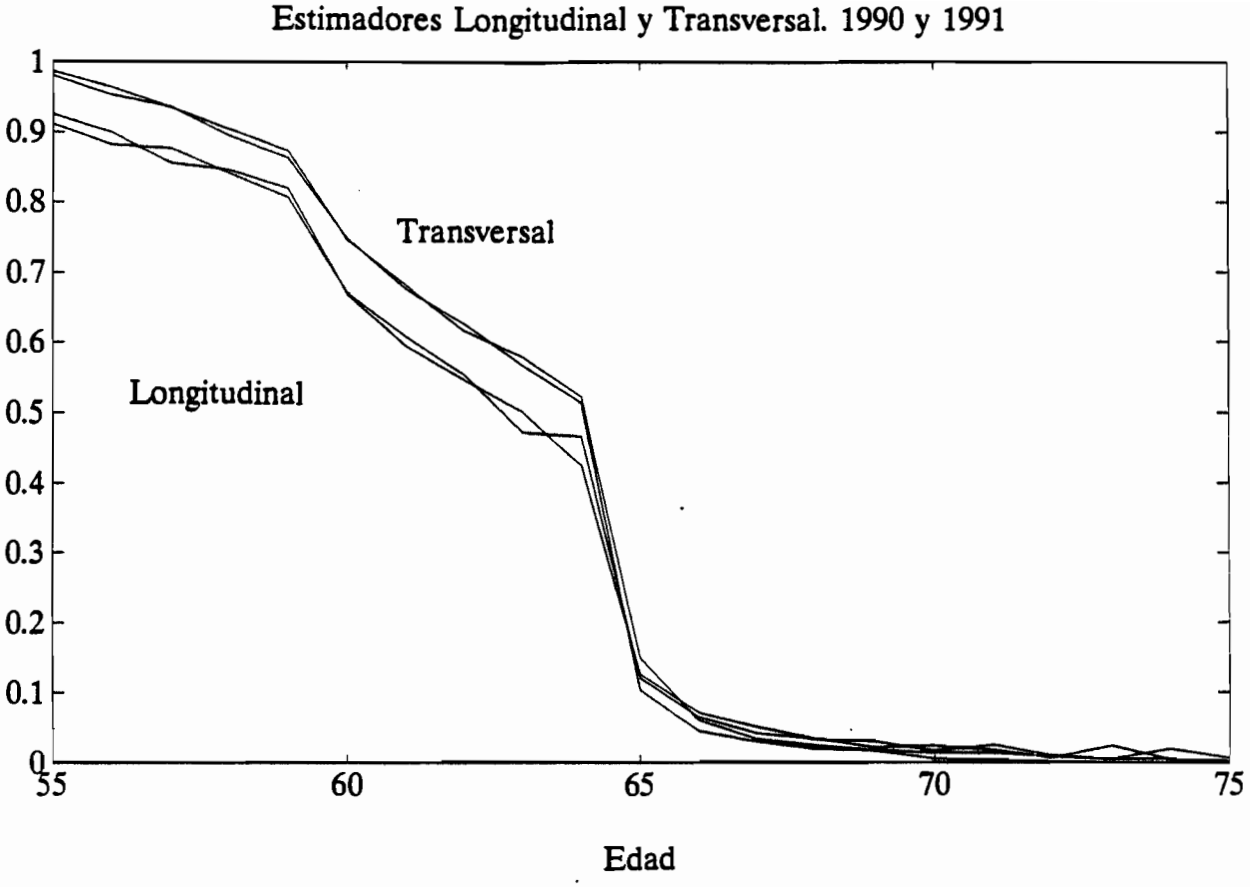
Estimadores Longitudinal y Transversal. 1990 y 1991
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
55
60
65
70
75
Transversal
Longitudinal
Edad

Figura 4

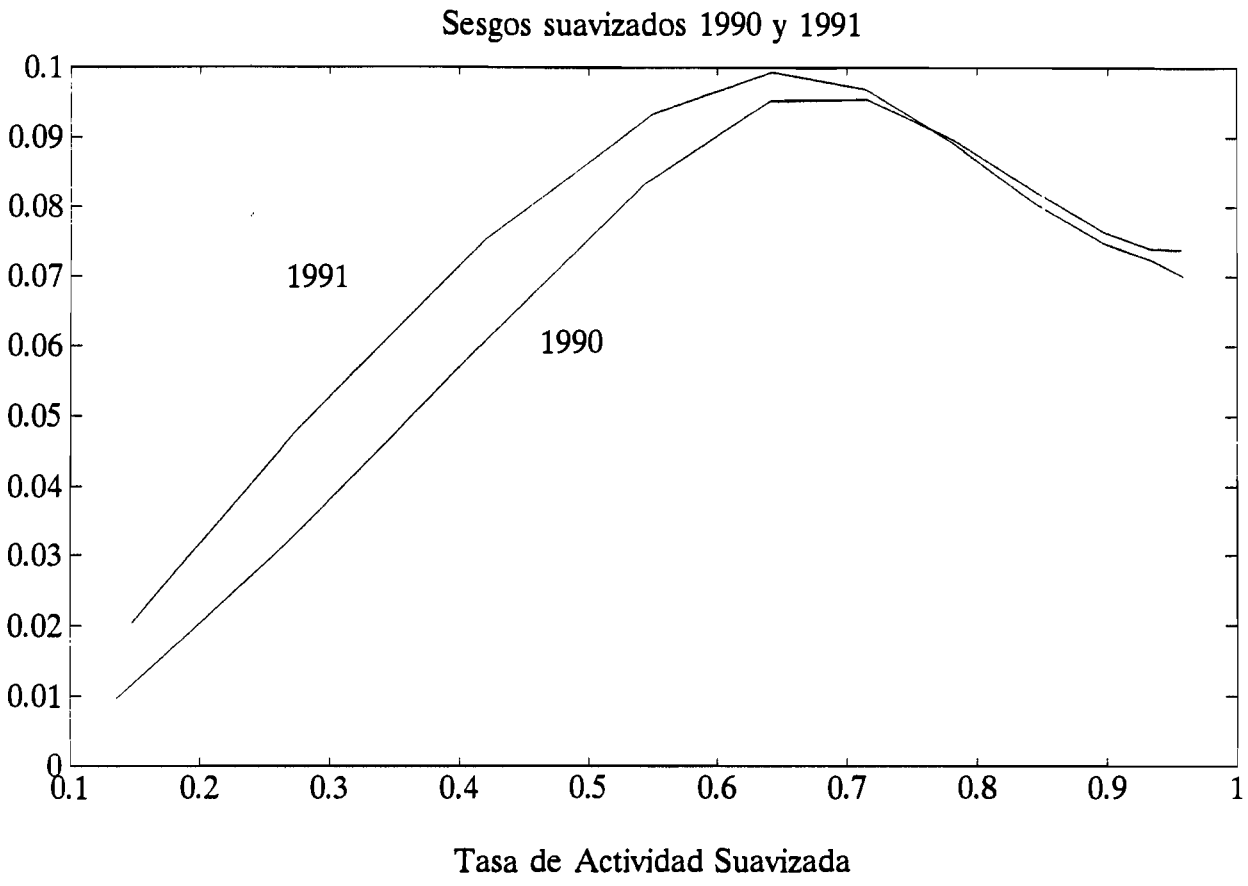
Sesgos suavizados 1990 y 1991
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1991
1990
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Tasa de Actividad Suavizada

Figura 5